基于布局图的多物体场景新视角图像生成网络 *

高小天[1a,1b]，张　乾[2]，吕　凡[2]，胡伏原[1a,1c†]

(1. 苏州科技大学 a. 电子与信息工程学院; b. 苏州市虚拟现实智能交互及应用技术重点实验室; c. 苏州市大数据与信息服务重点实验室，苏州 江苏 215009; 2. 天津大学智能与计算学部，天津 300354)

摘　要：新视角图像生成任务指通过多幅参考图像，生成场景新视角图像。然而多物体场景存在物体间遮挡，物体信息获取不全，导致生成的新视角场景图像存在伪影，错位问题。为解决该问题，提出一种借助场景布局图指导的新视角图像生成网络，并标注了全新的多物体场景数据集(Multi-Objects Novel View Synthesis，MONVS)。首先，将场景的多个布局图信息和对应的相机位姿信息输入到布局图预测模块，计算出新视角下的场景布局图信息；然后，利用场景中标注的物体边界框信息构建不同物体的对象集合，借助像素预测模块生成新视角场景下的各个物体信息；最后，将得到的新视角布局图和各个物体信息输入到场景生成器中构建新视角下的场景图像。在 MONVS 和 ShapeNet Cars 数据集上与最新的几种方法进行了比较，实验数据和可视化结果表明，在多物体场景的新视角图像生成中，所提方法在两个数据集上都有较好的效果表现，有效地解决了生成图像中存在伪影和多物体在场景中的位置信息不准确的问题。

关键词：多物体场景；遮挡现象；图像伪影；布局图；新视角图像生成

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2022.01.0032

Multi-object scenes novel view synthesis via layout projection

Gao Xiaotian[1a,1b],Zhang Qian[2],Lyu Fan[2],Hu Fuyuan[1a,1c†]

(1. *a. College of Electronic & Information Engineering*, *b. Suzhou Key Laboratory for Big Data & Information Service*, *c. Virtual Reality Key Laboratory of Intelligent Interaction & Application Technology of Suzhou*, *Suzhou University of Science & Technology*, *Suzhou Jiangsu* 215009, *China*; 2. *College of Intelligence & Computing*, *Tianjin University*, *Tianjin* 300354, *China*)

Abstract: The task of Novel View Synthesis refers to generating a new perspective image of the scene through multiple reference images. However, there are occlusions between objects in multi-object scenes, and object information cannot be fully obtained, resulting in artifacts and dislocation problems in the generated new-view scene images. In order to solve this problem, this paper proposes a new perspective image generation network guided by the scene layout map, and annotates a new multi-object scene dataset. (Multi-Objects Novel View Synthesis, MONVS) . First, input multiple layout information of the scene and the corresponding camera pose information into the layout prediction module, and calculate the layout information of the scene under a new perspective; Then, use the bounding box information of the objects marked in the scene to construct an object set of different objects, and use the pixel prediction module to generate the information of each object in the new perspective scene; Finally, input the obtained new perspective layout and various object information into the scene generator to construct a scene image under the new perspective. Compared with the latest methods on the MONVS and ShapeNet Cars data sets, Experimental data and visualization results show that in the new perspective image generation of multi-object scenes, the method in this paper has good performance on both data sets. Effectively solve the problem of artifacts in the generated image and inaccurate position information of multiple objects in the scene.

Key words: multi-object scene; occlusion; image artifacts; layout; novel view synthesis

0　引言

新视角图像生成(Novel View Synthesis, NVS)任务是在给定多幅输入图像和对应相机位姿情况下，生成物体或场景任意视角的图像。该任务在虚拟现实技术、机器人技术、静态图像动画制作等方面有着广泛的应用。因其避免了在生成任意视角图像过程中构建复杂三维模型，提升了生成效率，引起了学者的广泛关注。

早期的新视角图像生成方法是基于相机成像的相关知识，在像素空间或光线空间中利用插值的方法生成新视角图像[1]。随着深度学习的发展，文献[2]利用卷积网络生成刚性物体的新视角图像,但是仅利用卷积网络无法生成物体的细节信息且生成图像轮廓模糊；之后的工作将物体的先验知识加入模型训练中，取得较好的效果。以物体的几何先验[3,4]作为指导，将输入图像的像素值，根据物体的几何形状或 3D 点云信息[5~7]，投射到输出图像上。上述工作在单物体的新视角图像生成中取得了良好效果，但是，在更加真实的多物体场景中，由于将场景看做是一个整体[8]，当场景中物体间存在遮挡现象时，模型无法提取被遮挡物体的特征，也无法学习其几何信息，导致生成的图像出现模糊和伪影等错误，甚

收稿日期：2022-01-22；**修回日期：**2022-03-12　**基金项目：**国家自然科学基金资助项目(61876121)；江苏省重点研发计划项目(BE2017663)；江苏省教育厅高等学校自然科学研究面上项目(19KJB520054)

作者简介：高小天(1997-)，男，江苏徐州人，硕士研究生，主要研究方向为计算机视觉、深度学习和新视角图像生成；张乾(1991-)，男，硕士，主要研究方向为主动视觉、场景微变监测、光照重现；吕凡(1993-)，男，博士研究生，主要研究方向为连续学习、多模态学习、多任务学习；胡伏原(1978-)，男(通信作者)，教授，硕导，博士，主要研究方向为机器学习及计算机视觉(fuyuanhu@usts.edu.cn).

至会出现物体丢失的现象。该现象如图 1(a)所示。为了解决伪影问题，文献[9]利用深度图作为先验信息，指导网络生成场景的新视角图像，但深度图的获取需要精密的仪器，并且深度图无法改善由于多物体之间遮挡导致的边界模糊现象。

相较于深度图，包含图像中所有物体类别和边界框的布局图[10]，更容易获取。受布局图生成图像工作的启发，本文提出一种以场景布局图作为先验信息的新视角图像生成网络，如图 1(b)所示。与之前提出的其他基于深度学习的新视角图像生成方法相比，不需要获取复杂的场景深度图和点云信息，并且本文方法可以应用在多物体场景中。

首先根据场景中不同视角下的布局图信息，计算出场景中各个物体的旋转轨迹，通过目标相机和输入相机之间的位姿关系，得到新视角下图像的布局图，有效解决了由于遮挡导致生成图像中物体位置不准确的问题。基于布局图信息，对整个场景进行裁剪，将多物体新视角图像生成任务转换成多个单物体的新视角生成任务。为保证生成的单物体图像的细节完整，使用像素预测器使模型随着输入图像的变换逐步改善其生成结果。最后根据计算得到新视角布局图这一先验信息指导场景生成器生成整个场景的图像。

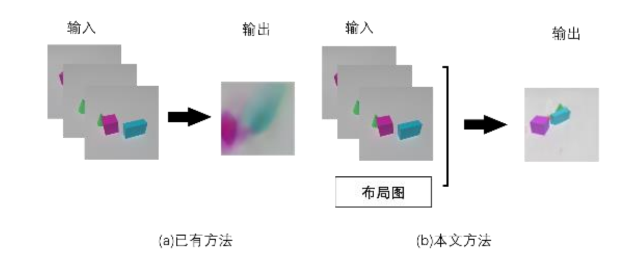


图 1 方法对比

Fig. 1 Method comparison

1 相关工作

1.1 布局图生成图像

使用附加信息[11,12](例如类别信息、文本描述[13]、场景图) 作为先验知识输入网络来指导网络生成图像是目前图像生成的主要做法之一。然而，如深度图等先验信息往往受各种条件的约束，难以获取。因此，一些工作引入了易获取的布局图作为先验知识。

在文献[10]中，布局图和对象信息被用于文本生成图像和场景图生成图像的任务中。通过将对象的形状与存储库中的特征进行匹配，从给定的布局图中生成新的场景图像。文献[14]提出了一种可以改变图像布局和对象风格的方法，通过改变布局图中边界框的大小和移动边界框，重构整个布局图，根据新构建的布局图信息生成图像。最新的工作，对布局图生成图像网络进一步优化[15,16]，通过从源视图中估计出整个场景的布局图，以深度信息作为约束，生成房间的平面图。

受到这种改变布局图生成图像模型的启发，考虑随着视角变化，物体的相对位置和边界框也发生变化。根据这种对应关系，从已知视角下的场景布局图推导新视角下的场景布局图，用来指导新视角图像的生成。

1.2 新视角图像生成

新视角图像生成是指通过给定多幅输入图像和相机位姿的情况下，生成物体或场景的任意视角下的新图像。

早先的一些工作中[3][4]，是基于物体的几何形状，将输入图像的像素通过映射或插值的方法，扭曲到新视角图像中，但是这种方法生成图像在细节纹理方面的渲染效果并不理想，并且无法生成源视图中缺失的像素；随着深度学习的发展，文献[2]通过卷积神经网络，根据源视图直接生成新视角图像，这种方法在单一刚性物体(如椅子，汽车等)的数据集上取得了不错的效果，但同样无法生成缺失的像素；为了解决像素缺失问题，Sun 等[17]提出光流预测模块和像素生成模块组成的网络，通过光流预测将输入图像中的像素映射到新视角图像中，像素预测模块根据输入图像生成缺失像素，以一种自学习置信聚合的机制生成新视角图像，但是这种方法对物体细节纹理的渲染仍不理想；随着深度图的发展，一些工作以图像的深度图[18]和图像的 3D 结构作为先验知识[3-6]，将源图像中的像素映射到目标图像中[19]；还有部分工作[20,21]通过重构场景或者物体的 3D 几何形状，再以新视角处的相机位姿为约束，生成图像，但是这种方法需要大量的时间和资源去进行训练；Mildenhall[22]等提出神经辐射场(Neural Radiance Fields, NeRF)这一全新的网络用来实现新视角图像的生成。该方法使用一个由空间三维坐标和观看方向组成的 5D 向量作为输入，输出物体上每个点的颜色和体积密度，在复杂场景中，取得了很好的效果。但是 NeRF 需要大量的输入视图来训练单个场景的模型，训练出的模型只能适用于单一的场景，泛化能力很差；Yu 等[23]对该方法进行了优化，提出 pixelNeRF 网络，能够使用少量的输入图像完成场景的重建，在训练时间和泛化性上取得了良好的进展，但依然无法解决场景中物体间遮挡导致的生成图像中存在伪影的问题。

2 布局图指导新视角图像生成方法

在本节中，介绍本文提出的基于布局图的多对象场景新视角图像生成方法。将布局图预测模块得到的新视角的布局图信息和像素预测器生成的各个物体的新视角图像输入场景生成器，生成新视角下的场景图像。整体架构如图 2 所示。

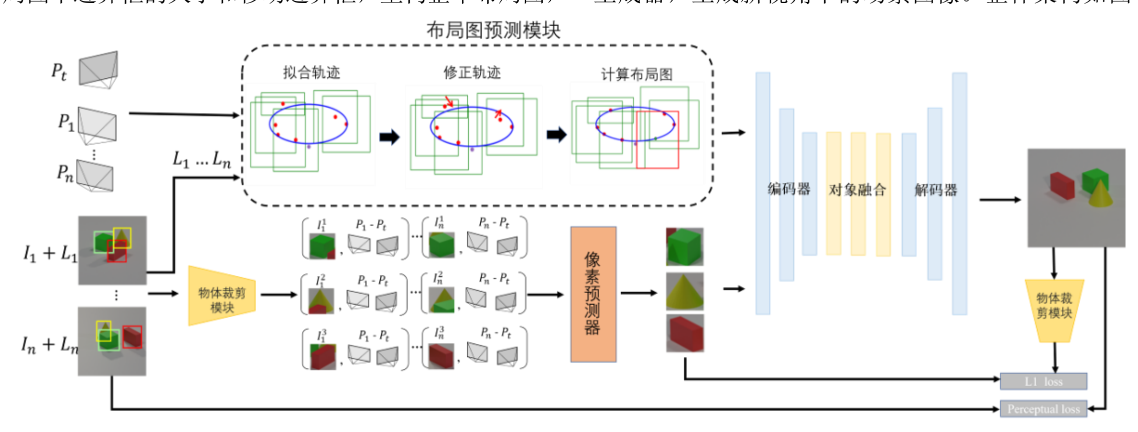


图 2 网络结构图

Fig. 2 Network structure diagram

输入多幅图像 I_n 及其相应的布局图 L_n，其中 $L_n=\{x_n^i,y_n^i,h_n^i,w_n^i\}$ 包括第 n 幅图像的中每个对象 O_i 的边界框信息(左上角坐标，高度，宽度)，将多个布局图 L_n 输入布局图预测模块，计算新视角下的布局图 L_t；模型对输入图像中的

每个对象实例 O_i 进行采样，再和相机位姿矩阵沿通道方向连接构建输入张量。将构建的张量输入像素预测器得到新视角下的各个物体的图像 I_t^i；最后，将 L_t 和 I_t^i 输入场景生成器中，物体图像 I_t^i 依次经过编码器和融合器，得到一个包含所有物体信息的融合特征，通过解码器生成场景图像。

2.1　布局图预测模块

借助相机标定[24~26]，将多幅输入图像中的物体映射到同一世界坐标系中，则同一物体在相机移动拍摄的过程中，可以看做是沿着一个椭圆的轨迹运动的。对单物体假设其初始轨迹椭圆 f 为：$Ax^2+By^2+Cxy+Dx+Ey+F=0$，其中 A，B，C，D，E，F 是椭圆的参数，利用 Faster-RCNN 目标检测方法得到输入图像的布局信息，可以通过多幅图像的布局图来计算上述椭圆的每个系数。

将输入图像对应的布局图输入布局图预测模块中，得到各个物体的运动轨迹并计算新视角下的布局图，布局图预测架构如图 3 所示。

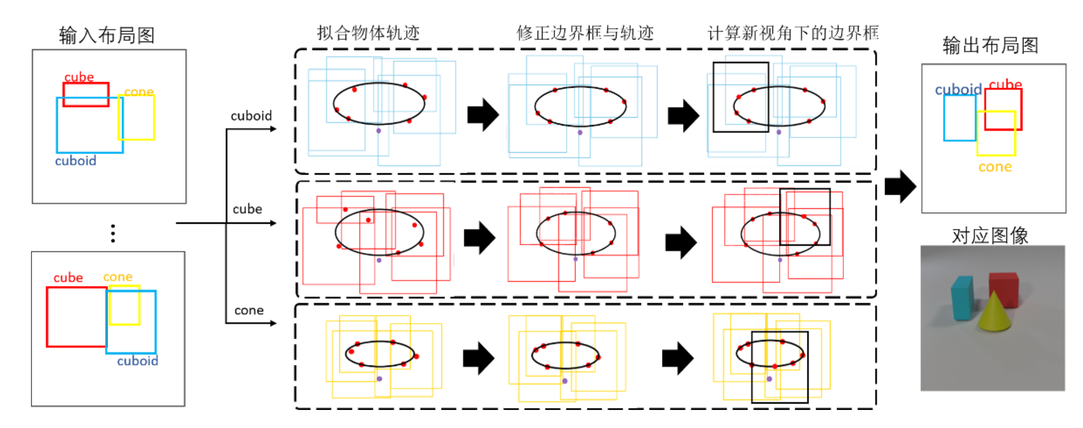


图 3　布局图预测框架

Fig. 3　Layout prediction framework

布局图信息按照物体类别 O_i 构建边界框集合 $L_n=\{x_n^i,y_n^i,h_n^i,w_n^i\}$，得到边界框中心坐标集合 $\left\{(x_1^i+\frac{1}{2}w_1^i,y_1^i+\frac{1}{2}h_1^i),(x_2^i+\frac{1}{2}w_2^i,y_2^i+\frac{1}{2}h_2^i),\ldots\ldots\right\}$，使用最小二乘法拟合公式曲线 f，求解椭圆参数 A，B，C，D，E，F。

由于场景中存在遮挡，使得标注出的物体边界框存在误差。为修正轨迹和物体边界框误差，提出一种迭代计算的方法进行边界框的修正和轨迹方程的优化。首先，计算边界框中心坐标和轨迹曲线 f 之间的最短距离 d，与设置的阈值比较，判定出需要修正的边界框中心坐标。需要修正的坐标每次以 $d/2$ 的步长向轨迹曲线逼近。然后，每一次更新后的坐标中心点，计算与上一次边界框四个顶点坐标的距离，以最大值为约束，对边界框进行扩充。得到更新后的边界框。最后，更新的坐标重复上述的操作，进行迭代训练，得到 $d_{\min}$ 最优解。目标函数 $d_{\min}$ 表示如下，

$$d_{\min}=\sqrt{(x_n^i+\frac{1}{2}w_n^i-f_x)^2+(y_n^i+\frac{1}{2}h_n^i-f_y)^2}\tag{1}$$

其中，f_x，f_y 是椭圆轨迹上的点集合。

一般来说，同一物体的边界框大小和距相机距离呈线性关系。将拟合出的轨迹曲线分为左右两个部分。左右两个部分的椭圆轨迹上，物体中心坐标的 y 值与物体边界框的宽高分别呈规律分布，即对象坐标越接近椭圆的下半圆时，表示对象距离拍摄的位置越近，对象的边界框越大，反之，边界框越小。为了计算出场景在新视角下的布局图信息，将修正后的边界框与相机位姿信息通过坐标系转换的方法，构建两者之间的关系，表示如下，

$$\begin{cases}y=k_1w+b_1\\y=k_2h+b_2\end{cases}\tag{2}$$

解出其参数 k_1，b_1，k_2，b_2，利用新视角处的相机位姿对应坐标计算出新视角下物体对应的边界框。

2.2　基于布局图生成新视角图像

2.2.1 像素预测器

现有的布局图生成图像方法通常通过卷积网络提取特征的方式生成图像，但是这种方法往往只关注图像纹理的转移，图像的细节和物体的几何形状无法完整的保留下来。为了解决这个问题，本文引入一个像素预测器，通过直接回归像素值，从源图像中预测目标图像中缺失的像素，保留了场景中各个物体的细节纹理，通过布局图中包含的对象类别信息对物体的几何形状进行约束，使得生成图像的结构保持一致。它是一种编码器-解码器类型的网络，在瓶颈层中使用卷积长短时记忆模块(Convolutional Long-Short-Term Memory，ConvLSTM)，将卷积层中提取到的信息通过 ConvLSTM 传递到对应的反卷积层中，使得获取的信息更丰富。

多视角输入图像通过像素预测器各自生成新视角下的图像，再将所有的图像以均值聚合，最终生成目标图像。细节如图 4 所示。首先使用独热编码(One-Hot)将输入视角的离散相机位姿进行矢量化处理，根据拍摄场景的相机总数 n，编码成 n 维元素的矢量，计算输入视角当前相机位姿 P_s 和目标位姿 P_t 之间的差值 P_{diff}；将 P_{diff} 输入网络，沿空间维度平铺 P_{diff} 获得输入的位姿张量 $P_{input}\in\mathbb{R}^{H\times W\times v}$，其中 v 表示位姿向量的维度。然后，根据获取的边界框 L_N 对输入图像进行裁剪，根据物体类别获取 i 组图像 I_s^i，对其进行双线性插值与位姿张量 P_{input} 沿着通道数连接，最后输入到像素预测器。像素预测过程可以表示为

$$I_P^i=P(I_s^i\circ P_{input})\tag{3}$$

其中，$P(\cdot)$ 表示像素预测器，I_P^i 是输入图像的预测图像，$\circ$ 表示沿着通道方向进行 Concat 操作。预测结果如图 5 所示，ShapeNet 数据集通过基于特征的方法得到的预测结果，只能生成汽车的轮廓，却无法保留汽车的细节纹理；而通过基于像素的方法得到的预测结果，汽车的细节纹理也被完整的生成。

I_{target} 通过所有预测图像 I_P^i 聚合生成，像素生成器被训练成最小化以下等式，

$$L_p=\frac{1}{N}\sum_{t=0}^{n}\left\|I_{\text{target}}-I_P^i\right\|_1\tag{4}$$

2.2.2 场景生成器

将经过像素预测器后，预测生成的物体图像与边界框 L_N 构建对象特征图 F_i，输入场景生成器生成新视角下的场景图像。对象类别 y_i 首先通过 Word Embedding 进行编码，然后，将类别编码 y_i 与对象特征 Z_i 串联起来，填充在对象边界框 L_N 内，

$$F_i=L_N\otimes(y_i\oplus Z_i)\tag{5}$$

其中，⊕ 表示矢量连接算，⊗ 表示将对象信息复制到边界框内。

为了将所有对象实例编码在期望的位置，在场景生成器中的解码器之后，加入一个多层卷积长短时记忆网络用来融合采样得到的对象特征，最终输出一个融合后的隐藏布局图 H ,其中包含所有对象的位置，类别和特征信息。隐藏布局图 H 输入解码器生成目标图像。

为了引导场景生成器中编码器、对象融合器、解码器能够合成真实的图像，防止融合生成的隐藏布局图 H 出现特征丢失。使用相同边界框 L_N 来裁剪生成的图像 I_{gen} 得到单个物体图像 I_s'，将 I_s' 输入到潜在代码估计器，获得物体的估计平均值和方差向量，然后直接使用计算出的平均向量作为回归的潜在代码 z_s'，并将其与像素预测器输出的值 Z_{si} 进行比较。具体表示为

$$L_1=\sum_{i=1}^{n}\|Z_{si}-Z'_{si}\|_1 \tag{6}$$

像素预测器生成的各个物体图像在场景生成器的融合过程中，由于边界框之间的重叠(现实场景中的遮挡)导致最终生成的场景图像中存在伪影问题。为了解决这个问题，本文采用基于 VGG-19 网络的感知损失(Perceptual Loss)[27,28]，感知损失定义如下：

$$L_{percept}(I_s,I_{gen})=\frac{1}{C_jH_jW_j}\|\phi_j(I_s)-\phi_j(I_{gen})\|_2^2 \tag{7}$$

其中，j 是 VGG-19 的中间层代号，本文使用的是 VGG-19 网络的 0，2，3 层提取的特征，$\phi_j(\cdot)$ 表示输入图像经过 VGG-19 的 j 中间层的输出。$C_jH_jW_j$ 是 $\phi_j(\cdot)$ 的通道，宽，高。

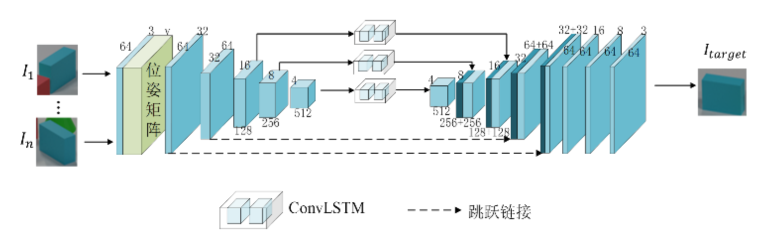


图 4　像素预测器

Fig. 4　Pixel predictor

3　实验结果分析

本文实验使用 PyTorch 深度学习框架，实验环境为 Ubuntu16.04 操作系统，使用 4 块 NVIDIA 1080Ti 的图像处理器(GPU)加速运算。



图 5　ShapeNet 数据集像素预测结果

Fig. 5　Shapenet dataset pixel prediction results

3.1　实验数据集

为了满足多物体场景下新视角图像生成任务的要求，构建两个不同难度的数据集。一个数据集是拍摄、标注的全新的数据集(Muliti Objects Novel View Synthesis Blender/Real, MONVS Blender/Real)。另一个数据集由 ShapeNet 中的对象合成多物体场景。

MONVS Blender/Real 数据集包含两部分，一部分为 MONVS Blender，另一部分为 MONVS Real。MONVS Blender 数据集包含不同类别的几何体，从 10 种颜色中随机抽取渲染物体且物体位置随机分布；MONVS Real 数据集中从 10 个不同类别的真实物体中随机抽取 3 个，单一颜色板作为背景；第二个数据集由 ShapeNet 中的对象合成，从 10 种不同的车型种随机抽取 3 辆车，双色板作为背景。选取 10 个位置放置相机，以固定的仰角获取场景图像。每个数据集各 100 个场景，包含 1000 幅图像。所有图像的分辨率均为 64×64。数据集中的随机样本图像如图 6 所示。

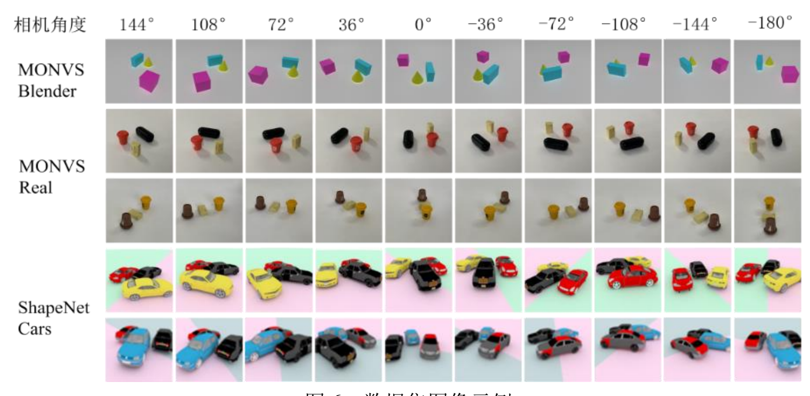


图 6　数据集图像示例

Fig. 6　Data set image example

3.2　实验结果分析

本文采用常用的结构相似性(SSIM)、峰值信噪比(PSNR)和感知相似度(LPIPS)对生成图像进行质量评估用以定量分析。

感知相似度(LPIPS)是近几年提出的一个新的图像评价指标，用于度量两张图像之间的差别。该度量标准学习生成图像到 Ground Truth 的反向映射，强制生成器学习从假图像中重构真实图像的反向映射，并优先处理它们之间的感知相似度。LPIPS 比传统方法(比如 L2/PSNR, SSIM, FSIM)更符合人类的感知情况。LPIPS 的值越低表示两张图像越相似，反之，则差异越大。计算公式如下：

$$d(x,x_0)=\sum_l\frac{1}{H_lW_l}\sum_{h,w}\|w_l\odot(\hat{y}_{hw}^l-\hat{y}_{0hw}^l)\|_2^2 \tag{8}$$

具体过程为：首先，训练过程中，将真实图像 x 和生成图像 x_0 送入神经网络(一般为训练好的 VGG19 模型)中进行特征提取，对第 l 层的特征进行激活后归一化处理得到 $\hat{y}^l,\hat{y}_0^l\in R^{H_l\times W_l\times C_l}$；然后，利用向量 $w_l\in R^{C_l}$ 缩放激活通道并计算

L_2 距离(w_i 是训练权重参数)，最后，在空间上求平均值，在通道上求和。

采用时间复杂度衡量每种方法复杂度的指标，即通过计算模型的浮点运算量(Floating-point Operations，FLOPs)。FLOPs 值越大，模型越复杂，反之模型越简单。

由于之前的工作没有与本文工作相同的设置，因此在对比方法的网络训练中只提供多视角的图像和相机位姿。经过测试，当输入图像为 6 幅时，轨迹方程拟合的准确率和时间为最佳。在三个数据集上，实验了 64×64 的图像分辨率，对于每个数据集各随机选择 800 幅图像用作训练，200 幅用作测试。给出了本文方法和最新的使用多视图生成新视角的方法 TB-network[29]、uORF-main[30]和 SVNVS[4]的定量结果比较。

图 7~9 中展示了本文的方法和其他最新的使用多视角图像作为输入的 NVS 的方法的可视化结果。这些结果涉及到在多物体场景中进行大视角转换的几个具有挑战性的示例。uORF-main 方法将单物体的 3D 表示和深度推理网络相结合，通过隐式搭建三维模型的方法实现新视角图生成任务，uORF-main 很难从输入图像中推测出目标视图中各个物体的对应关系。如图 7~9 结果中的第 3，4 列，uORF-main 应用在多物体场景中时，无法生成场景中物体的清晰图像；TB-network 方法通过网络首先生成高质量的 3D 结构和物体的体素信息，利用三维重建的方法生成新视角图像，但是 TB-network 无法较好的生成背景信息，因此许多背景细节丢失，容易产生空洞。例如图 7~9 结果中的第 5，6 列，生成图像中的各物体之间的背景产生空洞。SVNVS 通过输入图像以自监督的方式获取深度概率密度估计，来指导网络生成新视角图像，但是在处理大视角转换时，物体变化较大，目标图像的深度图无法通过输入图像的深度图准确生成，导致其网络生成图像中物体与物体之间的边界不清晰。例如图 7~9 结果中的第 7、8 列，当目标视角与输入视角相差过大时，各个物体的边界模糊，无法生成准确的图像。

相比之下，本文的方法通过场景的布局图信息指导网络生成新视角下的场景图像，不需要搭建场景的三维结构，不依赖输入图像的深度图，可以很好的恢复物体与物体和物体与背景之间的关系，生成的图像更加清晰真实。首先，本文方法在布局图信息的约束下，生成的图像中各个物体的形状和颜色相对清晰，其次，引入感知损失，使得物体与背景之间没有因为视角的转换而生成空洞和伪影。为了进一步证明本文的方法在多物体存在的场景下生成图像的真实性，对图 7~9 中生成的结果进行定性分析，结果如表 1 所列。当输入图像的个数相同时，本文的方法与其他多视图生成新视角方法在多物体场景数据集上的结构相似性(SSIM),峰值信噪比(PSNR)和感知相似度(LPIPS)的结果都是最好的。在 FLOPs 对比上，本文相较于其他三种模型有明显提升，这是由于本文提出的网络不需要进行 3D 信息的估计(如深度图和体素)，只需要对各物体的边界框进行计算就可以获取场景中各个物体的位置。这些结果表明，在相同情况下，布局图作为先验信息指导网络去生成新视角图像要优于使用深度概率密度估计和隐式三维结构的方案。

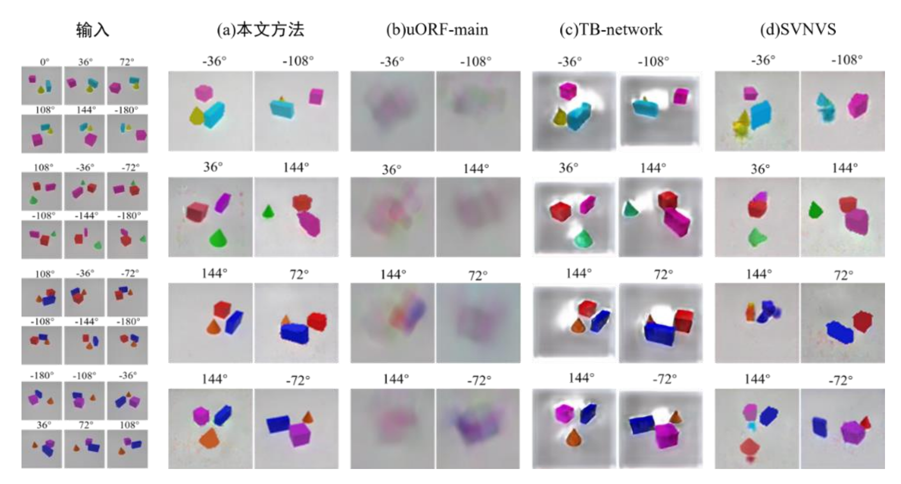


图 7　MONVS Blender 数据集对比实验结果

Fig. 7　MONVS Blender data set comparison experiment results

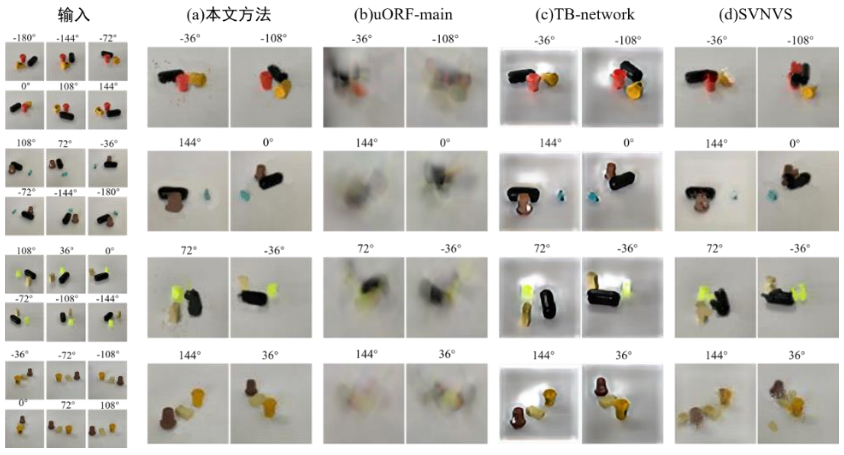


图 8　MONVS Real 数据集对比实验结果

Fig. 8　MONVS Real data set comparison experiment results

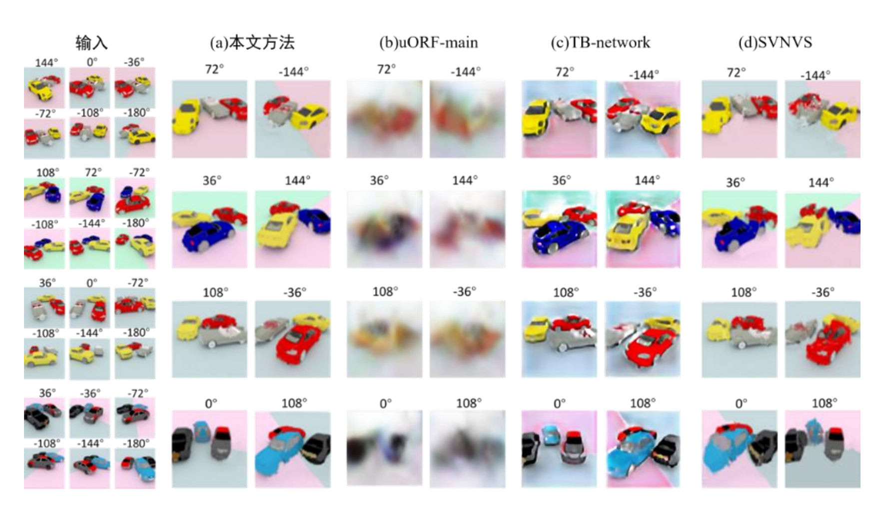


图 9　ShapeNet Cars 数据集对比实验结果

Fig. 9　Shapenet Cars data set comparison experiment results

表 1　数据集定量结果

Tab. 1　Data set quantitative results

方法	Blender			Real			ShapeNet Cars			
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	FLOPs/M
SVNVS[4](2021)	0.636	27.644	0.363	0.614	25.168	0.341	0.671	26.511	0.376	3577.426
TB-network[29](2019)	0.711	28.424	0.327	0.565	25.301	0.388	0.748	27.649	0.366	4832.501
uORF-main[30](2020)	0.474	28.041	0.493	0.391	24.087	0.471	0.492	27.241	0.454	2704.351
本文	**0.783**	**31.640**	**0.287**	**0.702**	**28.398**	**0.293**	**0.794**	**30.367**	**0.295**	**2402.218**

3.3　消融实验

为了验证所提模型中各个模块的有效性，在 MONVS 数据集上进行消融实验。可视化结果如图 10 所示，在没有预测布局图误差修正的情况下训练，生成图像中各个物体的位置不准确，图 10 第 2 行第 4 列所示,生成的长方体位置和圆锥位置与真实图像存在误差。这表明布局图预测模块能够准确的修正布局图信息，指导网络生成新视角图像。在没有引入感知损失的情况下，场景生成器生成的图像，存在严重的伪影现象，图 10 第 3 行第 2 列所示，在没有布局图作为先验信息的情况下，生成的圆锥体的位置和真实图像存在误差，并且生成的圆锥体不完整，出现像素丢失的现象。引入的感知损失保证了每幅图像中物体的周围没有出现伪影。图 10 第 4 行第 4 列所示，生成图像中长方体周围存在伪影。这是由于像素预测器生成的各个物体图像在场景生成器的融合过程中边界框之间的重叠(现实场景中的遮挡)导致的。这表明感知损失对图像的生成有严格的约束，并且有效解决了生成图像中的伪影问题。

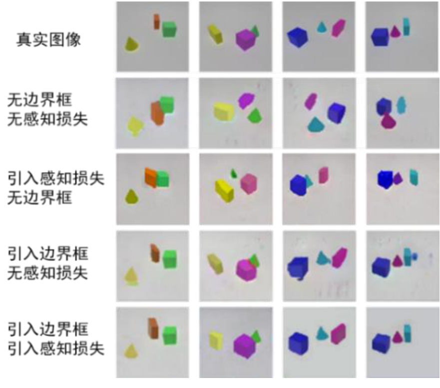


图 10　消融实验结果

Fig. 10　Results of ablation experiments

本文采用 FID 和 LPIPS 两个指标对消融实验的生成结果进行量化评估。定量结果如表 2 所列，将布局图预测模块修正后的场景布局图作为先验信息指导网络生成的图像，生成图像的真实度提高了 7.9%，证明了引入布局图可以有效解决场景中的物体在视角转换时，发生位置偏移的问题。模型加入感知损失后，生成图像的真实度提高了 58%，生成图像的颜色准确清晰，解决了生成图像中存在的伪影问题，证明了感知损失在提升图像质量上的有效性。

表 2　消融实验定量结果

Tab. 2　Quantitative results of ablation experiments

边界框	感知损失	FID↓	LPIPS↓
X	X	260.7	0.501
√	X	241.5	0.316
X	√	218.1	0.307
√	√	152.6	0.287

4　结束语

本文提出了一种以场景布局图为先验信息，指导网络实现场景新视角图像生成的方法。通过不同输入视角下的场景布局图信息，计算出新视角下的场景布局图，用来指导网络生成图像。解决了由于视角变换导致场景中物体丢失的问题，在场景生成器中加入感知损失函数，解决生成的各个物体在根据布局图信息进行集合时产生的伪影问题。实验结果表明，本文方法在多物体的简单场景下的新视角图像生成的性能和图像质量优于最近的几年的方法。然而，本文方法也有一些局限性，首先所提出的模型只能在环拍数据中对场景进行布局图预测；其次，对于新视角图像中前景与背景的交界处像素模糊。未来的工作将利用一些神经辐射场的方法，提高模型在拍摄不规则的数据集上的泛化性，使模型可以应用在园林等复杂的户外场景中。

参考文献：

[1]　Zhou T, Tulsiani S, Sun W, *et al*. View synthesis by appearance flow [C]// European conference on computer vision. Springer, Cham, 2016: 286-

301.

[2] Tatarchenko M, Dosovitskiy A, Brox T. Multi-view 3d models from single images with a convolutional network [C]// European Conference on Computer Vision. Springer, Cham, 2016: 322-337.

[3] Hani N, Engin S, Chao J J, *et al.* Continuous object representation networks: novel view synthesis without target view supervision [J]. Advances in Neural Information Processing Systems, 2020, 33: 6086-6099.

[4] Shi Y, Li H, Yu X. Self-Supervised Visibility Learning for Novel View Synthesis [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9675-9684.

[5] Song Z, Chen W, Campbell D, *et al.* Deep Novel View Synthesis from Colored 3D Point Clouds [C]// European Conference on Computer Vision. Springer, Cham, 2020: 1-17.

[6] Le H A, Mensink T, Das P, *et al.* Novel view synthesis from single images via point cloud transformation [J]. arXiv preprint arXiv: 2009. 08321, 2020.

[7] Park E, Yang J, Yumer E, *et al.* Transformation-Grounded Image Generation Network for Novel 3D View Synthesis [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 702-711.

[8] Choi I, Gallo O, Troccoli A, *et al.* Extreme view synthesis [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7781-7790.

[9] Huang P H, Matzen K, Kopf J, *et al.* Deepmvs: Learning multi-view stereopsis [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2821-2830.

[10] Zhao B, Meng L, Yin W, *et al.* Image generation from layout [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8584-8593.

[11] Herzig R, Bar A, Xu H, *et al.* Learning canonical representations for scene graph to image genera-tion [C]// European Conference on Computer Vision. Springer, Cham, 2020: 210-227.

[12] 兰红，刘秦邑. 图注意力网络的场景图到图像生成模型 [J]. 中国图像图形学报, 2020, 25 (08): 1591-1603. (Lan Hong, Liu Qinyi. A scene graph-to-image generation model for graph attention networks [J]. Chinese Journal of Image Graphics, 2020, 25 (08): 1591-1603.)

[13] 兰红，陈子怡，刘秦邑. 基于 Transformer 实现文本导向的图像编辑 [J/OL]. 计算机应用研究: 1-6 [2022-03-09]. http://www. arocmag. com/article/02-2022-05-032. html (Lan Hong, Chen Ziyi, Liu Qinyi. Text-Oriented Image Editing Based on Transformer [J/OL]. Application Research of Computers: 1-6 [2022-03-09]. http://www. arocmag. com/article/02-2022-05-032. html)

[14] Sun W, Wu T. Image synthesis from reconfigurable layout and style [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 10531-10540.

[15] Xu J, Zheng J, Xu Y, *et al.* Layout-Guided Novel View Synthesis from a Single Indoor Panorama [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16438-16447.

[16] Zou C, Colburn A, Shan Q, *et al.* Layoutnet: Recon-structing the 3d room layout from a single rgb image [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2051-2059.

[17] Sun S H, Huh M, Liao Y H, *et al.* Multi-view to novel view: Synthesizing novel views with self-learned confidence [C]// Proceedings of the European Conference on Computer Vision. 2018: 155-171.

[18] Flynn J, Neulander I, Philbin J, *et al.* Deepstereo: Learning to predict new views from the world's imagery [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5515-5524.

[19] Azinović D, Martin-Brualla R, Goldman D B, *et al.* Neural RGB-D surface reconstruction [J]. arXiv preprint arXiv: 2104. 04532, 2021.

[20] Guo P, Bautista M A, Colburn A, *et al.* Fast and Explicit Neural View Synthesis [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 3791-3800.

[21] 卫星，李佳，孙晓，等. 基于混合生成对抗网络的多视角图像生成算法 [J]. 自动化学报, 2021, 47 (11): 2623-2636. (Wei Xing, Li Jia, Sun Xiao, *et al.* Multi-view image generation algorithm based on hybrid generative adversarial network [J]. Chinese Journal of Automation, 2021, 47 (11): 2623-2636.)

[22] Mildenhall B, Srinivasan P P, Tancik M, *et al.* Nerf: Representing scenes as neural radiance fields for view synthesis [C]// European conference on computer vision. Springer, Cham, 2020: 405-421.

[23] Yu A, Ye V, Tancik M, *et al.* pixelnerf: Neural radiance fields from one or few images [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4578-4587.

[24] Zhang J, Yu H, Deng H, *et al.* A robust and rapid camera calibration method by one captured image [J]. IEEE Trans on Instrumentation and Measurement, 2018, 68 (10): 4112-4121.

[25] 赵漫丹，刘烨斌，吴高昌，等. 强约束条件下环绕式相机标定方法 [J]. 计算机应用研究, 2017, 34 (11): 3463-3467. (Zhao Mandan, Liu Yebin, Wu Gaochang, *et al.* Surround camera calibration method under strong constraints [J]. Application Research of Computers, 2017, 34 (11): 3463-3467.)

[26] 汪蕾，刘涛，董琦聪，等. 散焦模糊量估计的相机加权标定方法 [J]. 计算机辅助设计与图形学学报, 2020, 32 (3): 410-417. (Wang Lei, Liu Tao, Dong Qicong, *et al.* Camera weighted calibration method for defocus blur estimation [J]. Journal of Computer Aided Design and Graphics, 2020, 32 (3): 410-417.)

[27] Yang Q, Yan P, Zhang Y, *et al.* Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss [J]. IEEE Trans on medical imaging, 2018, 37 (6): 1348-1357.

[28] 吴从中，陈曦，季栋，等. 结合深度残差学习和感知损失的图像去噪 [J]. 中国图像图形学报, 2018, 23 (10): 1483-1491. (Wu Congzhong, Chen Xi, Ji Dong, *et al.* Image Denoising Combined with Deep Residual Learning and Perceptual Loss [J]. Chinese Journal of Image Graphics, 2018, 23 (10): 1483-1491.)

[29] Olszewski K, Tulyakov S, Woodford O, *et al.* Transformable bottleneck networks [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7648-7657.

[30] Yu H X, Guibas L J, Wu J. Unsupervised discovery of object radiance fields [J]. arXiv preprint arXiv: 2107. 07905, 2021.